Toin University of Yokohama
Research Bulletin, No.42, June 2020

Articles

Improving the Accuracy of the kNN Method:
The Use of the Best Combination of Features

PALACIOS PAWLOVSKY Alberto*, SHIOYA Minami[1] and SHIMIZU Haruki[1]

(Received Date: March. 14, 2020)

I. Introduction

The advances in circuit integration now make possible the use of very powerful personal computers. These machines allow us to deal with some combinational problems that a few years ago required too long running times and were considered impractical. In this paper we show the best predictive results obtained with the kNN (k Nearest Neighbor) method[1)] targeting two medical data sets of the UCI (University of California at Irvine) repository[2, 3)]. Both are breast cancer datasets. The database of the university of Coimbra is one created to develop a prediction model of breast cancer, based on data and parameters gathered in routine blood analysis[4)]. The second database is one used for medical diagnosis applied to breast cytology[5)].

In the following section we detail these two datasets and how we used them for prediction of breast cancer.

In section III we describe how we used the kNN method for prediction and the results obtained. Section IV details the conclusions and some topics for further research.

II. Breast cancer datasets

The two datasets used in this work are related to breast cancer in women. The first one and one of the newest in the UCI repository was donated by a research group of the Faculty of Medicine of the University of Coimbra (abbreviated CBCD). They first used this dataset in [6].

The second dataset was donated by Dr. William H. Wolberg of the University of Wisconsin during three years from 1989 to 1991 (abbreviated WOBCD).

1. Coimbra dataset

The Coimbra dataset has data of 116 patients. Of the 116 patients, 62 women have breast cancer and 52 are healthy ones. Each patient datum includes age (years), BMI (kg/m^2), glucose (mg/dL), insulin (μU/mL), HOMA, Leptin (ng/mL), Adiponectin (μg/mL), Resistin (ng/mL), MCP-1 (pg/dL) and label (1 = healthy, 2 = with breast cancer). The nine features that can be used as predictors were collected by the same research physician and during the first consultation. The blood related

* PALACIOS PAWLOVSKY Alberto: Professor, Department of Clinical Engineering, Faculty of Biomedical Engineering, Toin University of Yokohama, 1614, Kurogane-cho, Aoba-ku, Yokohama 225–8503, Japan

[1] SHIOYA Minami and SHIMIZU Haruki: Students, Department of Clinical Engineering, Faculty of Biomedical Engineering, Toin University of Yokohama.

data was gathered in routine blood analysis.

2. Wisconsin breast cancer diagnosis dataset

The dataset available at the UCI repository contains data of 699 patients. We removed 16 instances due to missing values. Each instance contains 9 features. These are clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. It is a well-known dataset and is used frequently in data mining related research[7)].

III. Evaluation of the best combination of features using kNN

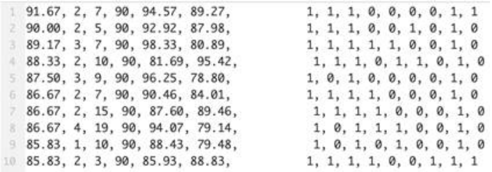
The kNN method is a well-known classification method in the area of machine learning, it is easy to implement and has good accuracy. It has been used for diagnosis and prognosis of several diseases[8–13)].

1. On the use of the kNN method

We used an in-house implementation of the kNN method that allows us to set many different parameters of it. We evaluated all the 511 combinations of features of the Coimbra and Wisconsin datasets. We evaluated all them using a kind of ten-fold cross validation. We run ten times the kNN method with the same setting to obtain the average accuracy with each combination. Our implementation uses six different metrics (distances). We used the Euclidean, Manhattan, Chebyshev, Sorensen, Canberra, and Mahalanobis distances. In our evaluations the kNN method uses nine different sizes of all the available data as data used for prediction. It uses 10%, 20%, 30%, …, 90% of all data to predict if a patient has breast cancer or not (a total of nine sizes). We also make the kNN do the prediction for all the possible values of k (the number of data (neighbors) close to the data we want to classify). We also preprocessed the data in two ways. The first one uses the minimum and maximum values of each feature to standardize the data (we call it normalization and show it in some figures and tables as 'nor'). The second one uses the mean value and standard deviation of each feature to standardize the data (we call it standardization and point to it as 'std'). To differentiate the UCI data from these two ones we abbreviate it as 'raw'.

2. Results for the Coimbra dataset

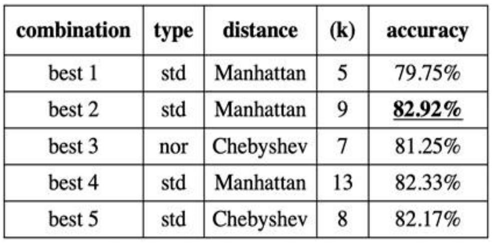
The results obtained running each setting using the kNN method gave us the top ten combinations given in ***Figure 1***. The features used by kNN are shown in this figure on the right side. When a feature is used it is signaled by a one and zero otherwise. The combination that gave us the best average accuracy of 91.67% used age, BMI, glucose, Resistin, and MCP-1.



```
91.67, 2, 7, 90, 94.57, 89.27,          1, 1, 1, 0, 0, 0, 0, 1, 1
90.00, 2, 5, 90, 92.92, 87.98,          1, 1, 1, 0, 0, 1, 0, 1, 0
89.17, 3, 7, 90, 98.33, 80.89,          1, 1, 1, 1, 1, 0, 0, 1, 0
88.33, 2, 10, 90, 81.69, 95.42,          1, 1, 1, 0, 1, 1, 0, 1, 0
87.50, 3, 9, 90, 96.25, 78.80,          1, 0, 1, 0, 0, 0, 0, 1, 0
86.67, 2, 7, 90, 90.46, 84.01,          1, 1, 1, 1, 0, 0, 0, 1, 0
86.67, 2, 15, 90, 87.60, 89.46,          1, 1, 1, 1, 0, 0, 0, 1, 0
86.67, 4, 19, 90, 94.07, 79.14,          1, 0, 1, 1, 1, 0, 0, 1, 0
85.83, 1, 10, 90, 88.43, 79.48,          1, 0, 1, 0, 1, 0, 0, 1, 0
85.83, 2, 3, 90, 85.93, 88.83,          1, 1, 1, 1, 0, 0, 1, 1, 1
```

Fig. 1 Best 10 combinations for CBCD.

To obtain a more accurate value of the average accuracy that we can expect using kNN, we evaluated the first five combinations shown in ***Figure 1*** running them 100 times. The corresponding average accuracy results are shown in ***Figure 2***.



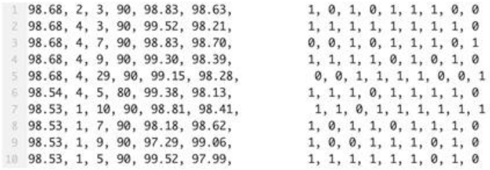
combination	type	distance	(k)	accuracy
best 1	std	Manhattan	5	79.75%
best 2	std	Manhattan	9	**82.92%**
best 3	nor	Chebyshev	7	81.25%
best 4	std	Manhattan	13	82.33%
best 5	std	Chebyshev	8	82.17%

Fig. 2 Best 5 combinations results: CBCD.

We can note that the best combination of features (best 1, first one of ***Figure 1***) has the smallest average accuracy. One more characteristic that is worth to notice is that all other combinations (best 2 ~ best 5) do use Resistin, but do not use (the last feature) MCP-1. Moreover, the combination with the highest average accuracy uses almost the same features reported in [4].

3. Results for the Wisconsin dataset

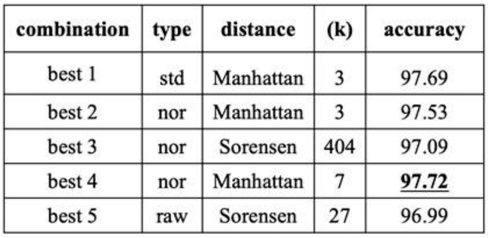
The results obtained with the Wisconsin dataset are shown in ***Figure 3***. The top five combinations gave us the best average accuracy of 98.68% when running each one for 10 times with each possible setting.



```
98.68, 2, 3, 90, 98.83, 98.63,          1, 0, 1, 0, 1, 1, 1, 0, 0
98.68, 4, 3, 90, 99.52, 98.21,          1, 1, 1, 1, 1, 1, 1, 1, 0
98.68, 4, 7, 90, 98.83, 98.70,          0, 0, 1, 0, 1, 1, 1, 0, 1
98.68, 4, 9, 90, 99.30, 98.39,          1, 1, 1, 1, 0, 1, 0, 1, 0
98.68, 4, 29, 90, 99.15, 98.28,          0, 0, 1, 1, 1, 1, 0, 0, 1
98.54, 4, 5, 80, 99.38, 98.13,          1, 1, 1, 0, 1, 1, 1, 1, 0
98.53, 1, 10, 90, 98.81, 98.41,          1, 1, 0, 1, 1, 1, 1, 1, 1
98.53, 1, 7, 90, 98.18, 98.62,          1, 0, 1, 1, 0, 1, 1, 1, 0
98.53, 1, 9, 90, 97.29, 99.06,          1, 0, 0, 1, 1, 1, 0, 1, 0
98.53, 1, 5, 90, 99.52, 97.99,          1, 1, 1, 1, 1, 1, 0, 1, 0
```

Fig. 3 Best 10 combinations for WOBCD.

To obtain more accurate values of the average accuracy, we evaluated the first five combinations running them 100 times. The corresponding average accuracy results are shown in ***Figure 4***.



combination	type	distance	(k)	accuracy
best 1	std	Manhattan	3	97.69
best 2	nor	Manhattan	3	97.53
best 3	nor	Sorensen	404	97.09
best 4	nor	Manhattan	7	**97.72**
best 5	raw	Sorensen	27	96.99

Fig. 4 Best 5 combinations results: WOBCD.

The combination of features with the highest average accuracy was the fourth one of ***Figure 3***. It used the clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, bare nuclei, and normal nucleoli.

IV. Conclusions

The evaluation of all possible combinations of the features in the Coimbra data set gave, as best combination, one that uses almost the same features found to be the best in [4]. The best combination of [4] uses age, BMI, glucose, and Resistin. Our best combination adds to them leptin. This could hint that one more possible good (or best) combination is the one found by us. Additional work would be needed to confirm it.

Regarding the Wisconsin dataset we found that using clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, bare nuclei, and normal nucleoli would give a high average accuracy. The combination that we found is better than the one of [12] that used the clump thickness, single epithelial cell size, the bare nuclei, the bland chromatin, and the mitoses values and gave a best average accuracy of 97.4%.

We can see from the above results that the data must be processed for both datasets. The data of the Coimbra dataset must be standardized and the Wisconsin dataset must be normalized to obtain the best results. The best metric (distance) for both datasets seems to be the Manhattan distance.

Having found that the results of combinations, that are not the best when evaluating them for a running of 10 trials, could surpass the best one, it would be worth to evaluate the remaining best 10 combinations (best 6 ~ best 10) of ***Figure 3*** for the Wisconsin dataset.

[References]

1) E. Fix and J. L. Hodges, Jr., “Discriminatory analysis, nonparametric discrimination,” USAF School of Aviation Medicine, Randolph Field, Tex., Project 21–49–004, Rept. 4, Contract AF41(128)–31, February 1951.
2) https://archive.ics.uci.edu/ml/datasets/Breast+Can-

cer+Coimbra.

3) https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original).

4) M. Patrício, J. Pereira, J. Crisóstomo *et al*., "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," BMC Cancer 18, 29 (2018). https://doi.org/10.1186/s12885-017-3877-1.

5) W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," Proc. of the National Academy of Sciences of the USA, Vol.87 No.23, pp.9193–9196, 1990.
https://doi.org/10.1073/pnas.87.23.9193.

6) J. Crisóstomo, P. Matafome, D. Santos-Silva, et al., "Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer," Endocrine 53, pp.433–442, 2016.
https://doi.org/10.1007/s12020-016-0893-x

7) V. Kumar, B. K. Mishra, M. Mazzara, D. N. H. Thanh, A. Verma, "Prediction of Malignant & Benign Breast Cancer: A Data Mining Approach in Healthcare Applications," Advances in Data Science and Management: Proceedings of ICDSM 2019, pp.435–442.

8) A. Palacios Pawlovsky and M. Nagahashi, "A Method to Select a Good Setting for the kNN Algorithm when Using it for Breast Cancer Prognosis," Proceedings of the 2nd. IEEE International Conference on Biomedical and Health Informatics (BHI), pp.189–192, Sevilla, Spain, June 2014.

9) K. Odajima and A. Palacios Pawlovsky, "A Detailed Description of the Use of the kNN Method for Breast Cancer Diagnosis," Proc. of the 7th International Conference on Biomedical Engineering and Informatics (BMEI), pp. 606–610, Dalian, China, October 2014.

10) Alberto Palacios Pawlovsky, "A kNN Method that Uses a Non-natural Evolutionary Algorithm for Component Selection," International Symposium on Computational Intelligence & Applications ISCIA 2017, Malacca, Malaysia, July 14–15, 2017.

11) A. Palacios Pawlovsky, "An Ensemble Based on Distances for a kNN Method for Heart Disease Diagnosis," Proc. of the International Conference on Electronics, Information and Communication (ICEIC 2018), pp.6–9, Hawaii, USA, January 24–27, 2018.

12) Alberto Palacios Pawlovsky, "An Immune Algorithm with an Evolutionary Scheme for Component Selection for the kNN Method," Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2018), pp.2554–2560, IEEE World Congress on Computational Intelligence (WCCI 2018), Rio de Janeiro, Brasil, July 8–13, 2018.

13) Alberto Palacios Pawlovsky and Yuki Suzuki, "An Immune Algorithm that Uses a Master Cell for Component Selection for the kNN Method," Proceedings of the Global Medical Engineering Physics Exchanges and Pan American Health Care Exchanges GMEPE/ PAHCE 2019, pp.11–14, Buenos Aires, Argentine, March 26–31, 2019.